Nivard, M. G., Gage, S. H., Hottenga, J. J., van Beijsterveldt, C. E. M., Abdellaoui, A., Bartels, M., Baselmans, B. M. L., Ligthart, L., Pourcain, B. S., Boomsma, D. I., Munafò, M. R., & Middeldorp, C. M. (2017). Genetic Overlap Between Schizophrenia and Developmental Psychopathology: Longitudinal and Multivariate Polygenic Risk Prediction of Common Psychiatric Traits During Development. *Schizophrenia Bulletin*, *43*(6), 1197-1207. [sbx031]. https://doi.org/10.1093/schbul/sbx031

Peer reviewed version

Link to published version (if available):
10.1093/schbul/sbx031

Supplement

Figure notes and Table titles

Figure S1: Bubble plot of the relationship between the polygenicity prior and the effect size in the polygenic risk score analyses. Note the increase in effect size with the increase in polygenicity prior. The solid line indicates the best fit obtained from the meta-regression model (model 3). The dashed lines reflect the upper and lower confidence bounds.

Figure S2 Bubble plot of the approximated genetic correlations between schizophrenia and childhood psychopathology per disorder given the assumptions described in Supplementary Note 1. In this figure, we assume the variance explained by all markers in childhood psychopathology is constant and 20%. Circles indicate the transformed observed regression coefficients to genetic correlations (ALSPAC in red, NTR in blue). The size of the circles is proportional to the inverse of the variance, and thus larger circles reflect more accurate estimates. The solid line reflects the genetic correlation and the dashed lines indicate the upper and lower 95% confidence interval around the genetic correlation, quantifying the uncertainty in the meta-regression but not in the variance in childhood psychopathology explained by all measured markers, or the estimate of the number of independent markers.

Figure S3 Bubble plot of the approximated genetic correlations between schizophrenia and childhood psychopathology per disorder given the assumptions described in Supplementary Note 1. In this figure, we assume the variance explained by all markers in childhood psychopathology is constant and 10%. Circles indicate the transformed observed regression coefficients to genetic correlations (ALSPAC in red, NTR in blue). The size of the circles is proportional to the inverse of the variance, and thus larger circles reflect more accurate estimates. The solid line reflects the genetic correlation and the dashed lines indicate the upper and lower 95% confidence interval around the genetic correlation, quantifying the uncertainty in the meta-regression but not in the variance in

childhood psychopathology explained by all measured markers, or the estimate of the number of independent markers.

Table S1: Sample sizes per age group for the NTR, ALSPAC and combined

Table S2: Descriptives

Table S3: Prediction of non-participation based on the schizophrenia PRS

Table S4: Prediction of non-participation based on psychopathology scores at an earlier time point

Supplementary Note 1

This note accompanies the manuscript entitled: "Genetic overlap between schizophrenia and developmental psychopathology: a longitudinal analysis of common childhood disorders between age 7 and 15". All data described here and analyses presented here serve to support the conclusions of the manuscript as published.

Phenotype descriptives

Table S2 presents the mean scores on the DSM-IV based scales of anxiety, depression, ADHD, ODD/CD for males and females in the NTR at different ages (left) as well as the percentages of male and female ALSPAC participants with these diagnoses, defined as a score of 4 or 5 on the DAWBA (right). (Note that in the analyses, the 6-category DAWBA band was used as outcome variable since this is a more informative measure than the dichotomous DAWBA diagnosis).

Genotyping and genotype quality control:

The NTR participants were genotyped on Affymetrix 6.0, Affymetrix-perlegen 5.0, Illumina 660 and Omni express (1M) platforms. Array specific calls and cleaning were performed before data from different platforms were combined. Data from different platforms were strand aligned, SNPs with a minor allele frequency below 1%, a HWE p-value < $1*10^{-5}$ and with a genotype missingness rate >

10% or call rate < 95% were removed. Individuals with an excessive or low heterozygosity were removed (F > .10 or F < .10). After QC, genotypes were imputed to a common set of SNPs based on the goNL reference set.[1] SNPs were imputed that were not directly measured on each platform. Samples were excluded when reported gender did not match biological gender or when individuals were of non-European ancestry based on principle component analysis.[2] In the NTR, sex, call rate, F (inbreeding coefficient), five principle components based on global ancestry and five principal components correcting for local ancestry differences within the Netherlands were included as covariates in all analyses .[2]

In ALSPAC, children were genotyped on the Illumina HumanHap550 quad chip genotyping platforms. The raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches, minimal or excessive heterozygosity, disproportionate levels of individual missingness (>3%), and insufficient sample replication (IBD < 0.8). Population stratification was assessed by multidimensional scaling analysis, and compared with Hapmap II (release 22); all individuals of non-European ancestry were removed. SNPs with a minor allele frequency of < 1%, a call rate of < 95%, or evidence for violations of Hardy-Weinberg equilibrium ($p < 5E^{-7}$) were removed. Cryptic relatedness was measured as proportion of identity by descent (IBD > 0.1). Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation, though not association. Imputation of the target data was performed using Impute V2.2.2[3] against the 1000 genomes phase 1 version 3 reference panel, using all 2186 reference haplotypes (including non-Europeans).[4] As the ALSPAC sample, after QC, is assumed to be genetically homogeneous with respect to ancestry and local ancestry differences, no principal components were added as covariates, sex was included as covariate in all analyses.

The correction for the presence of overlapping subjects at the different ages of measurement, and the correlation between the polygenic predictors.

In the meta-regression analysis performed we had a set of 6 predictors X, and 32 outcomes Y. We performed a series of 192 univariate regressions:

$$Y_1 = X_1 B_1 + e$$

…

$$Y_{32} = X_6 B_{192} + e$$

We constructed an approximate error correlation matrix (i.e. the correlation between the regression parameters B) for a series of univariate regressions of p equal to:

$$cor(B) \approx (cor(Y)) \otimes (cor(X))$$

We specified the error *covariance* matrix as: $se * cor(B) * se$. Where *se* was a 192 x 192 diagonal matrix with the standard associated with each of the parameters B on the diagonal. The errors were assumed to be independent between cohorts and therefore correlations between cohorts were set to zero in matrix $cor(B)$. Based on the specified error correlation matrix, we performed the meta-analysis and meta-regression of the beta's obtained from the univariate regression analyses. To test whether the proposed error correlation matrix accurately accounted for the dependence induced by correlated predictors and outcomes, we performed type 1 error simulations.

We simulated 3 traits (Y) (correlations between .3 and .5) and 3 polygenic scores (X) (correlations between .9 and .8) for 100 subjects. In each simulation there was no true association between PRS and traits. We regressed each trait Y on each polygenic score X, and meta-analyzed the 9 test statistics obtained from these regressions, correcting for the dependence between traits and risk scores as outlined above. Given a small sample in the univariate regressions (N=100) the following

slightly liberal type 1 error rates were observed. The liberal type-1 error was likely induced by the fact that the test statistic obtained in each meta-analysis followed a t-distribution and not a normal distribution.

Alpha	Type 1 error
0.10	0.123
0.05	0.065
0.01	0.015

Simulating data given a larger sample of 1000 subjects in the initial univariate regressions the following, accurate, type 1 error rates were observed:

Alpha	Type 1 error
0.10	0.102
0.05	0.052
0.01	0.01

In our study, the sample size for the individual univariate regressions to be meta-analyzed ranged between 1200 and 6000 thus we were satisfied with the results of the type-1 error simulations.

A different limitation was that the error covariance as specified here assumed total sample overlap, and the absence of any covariates. However, we did include covariates to control for population stratification and mean differences between male and female participants. These effects were assumed to be sufficiently small to allow our approximation to be valid. Strong covariate effects and substantial dropout would likely reduce power to detect an overall or age effect, and possibly increase type 1 error in some situations. As a form of sensitivity analysis the off diagonal elements of

the phenotypic correlation matrices in ALSPAC and NTR were shrunk by 50% or 33% and increased by up to 10% to simulate the effect of less than total sample overlap or the effect of covariates changing the error covariance matrix. The conclusions remained virtually unchanged. Model 3 as described in the main text, had the best model fit when the off-diagonal elements in the phenotypic covariance matrix were reduced 50% or 33% and model 4 performed best when the phenotypic covariance was increased 10%. Parameter estimates and test statistics in model 3, fitted on the increased or decreased error covariance matrix were virtually unchanged. To conclude, the sensitivity analyses revealed that the effects of misspecification of the error covariance matrix probably did not influence the conclusions.

Parametric resampling of the data to account for the influence of sampling fluctuation

To quantify the influence of sampling fluctuation on our model selection, we resampled the input for the meta-regression from a multivariate normal distribution with means equal to the observed regression coefficients in the univariate PRS analyses, and covariance equal to the above specified error covariance matrix. Unlike non-parametric bootstrapping this technique makes assumptions about the asymptotic distributions of test statistics. However parametric resampling does allow for quantification of sampling variance in the model selection procedure.

We resampled 1000 datasets, on each of these the model selection procedure was repeated, the percentages in Table 1 (main text) reflect the percentage of resample datasets for which each model best fitted the data. The results represent the expected sample fluctuation in the model selection procedure induced by sample fluctuation, assuming the estimated effect sizes and error covariance are representative of the true effects and error covariance.

Mixed effects meta-regression to account for residual heterogeneity

In some cases a random effect which is unjustly omitted from the model can induce false positive results in meta-regression. Given only two cohorts are in the current study, a full mixed meta-regression makes little sense. However as a robustness check we fit two meta-regression models which allow for random effects and determine their influence on the results.

The best fitting fixed effects meta-analysis model (model 3; Table 1) revealed a moderate amount of residual variation not accounted for by the meta-regression model (Q_e = 226,49, df=181, p = 0.0122). We therefore performed two additional random effects meta-analyses. Our first random effect model allowed for (correlated) random effects for each observed effect size, where the correlations between the random effects were assumed to be equal to the error covariance. The first mixed effects model significantly improved the model fit (LRT=8.81, df = 1, p = 0.0009). The fixed effects age, ODD/CD, and agexADHD were all significant ($p < 0.05$) in this mixed effects model (as they were in the fixed effects meta-regression model), but the effect agexODD/CD no longer reached significance (p = 0.0681) . The omnibus test including all meta-regression parameters also remained significant (QM = 37.0610, df = 10, $p < 0.0001$). The second mixed effects model included a random intercept, i.e., in this model, the only dependence between effect sizes was introduced by the meta-regressors or the error covariance. This second random effects model also significantly improved model fit over the fixed effects model (LRT = 14.7982, df=1, $p < 0.0001$). The second mixed effects model revealed a significant age effect and agexADHD effect ($p < 0.05$), but no significant ODD/CD and agexODD/CD effects. The overall test of parameters remained significant (QM= 30.0032, df=10, p = 0.0009). To conclude, both random effects models retained the main conclusions as the fixed effects model, i.e., an increasing association between schizophrenia PRS and childhood psychopathology with age, and some differences between the disorders in their relationship with schizophrenia.

Estimating genetic correlations based on the results from the polygenic risk score analyses

The univariate polygenic risk analyses results were obtained from either an ordered logistic regression (ALSPAC) or general estimation equations (GEE in NTR). For explanatory purpose we consider an OLS regression:

$$\boldsymbol{y} = + * \boldsymbol{PRS} + \cdots + \boldsymbol{e}$$

where the trait (y) and the PRS are scaled to unit variance and centered. The regression contained a number of other covariates (such as sex and principal components). Assuming the effects of the principal components and sex on the phenotypes were small to negligible, the square of B_1 (B_1^2) is equal to the variance explained in the phenotype by the PRS. We further assume that the squared predicted outcome of the multivariate meta-analyses correspond to R^2. Given these assumptions we used the previously derived relationship between R^2 and genetic correlation[5] to approximate the genetic correlations between childhood psychopathology and schizophrenia:

$$R^2 = \sigma_{g1,g2} \frac{\sigma_{g1,g2} * \frac{N}{M}}{\frac{N}{M} * {\sigma_{g1}}^2 + 1}$$

The inverse relationship equals:

$${\sigma_{g1,g2}}^2 = R^2 \Big/ \frac{\frac{N}{M}}{\frac{N}{M} * {\sigma_{g1}}^2 + 1}$$

From which we can obtain:

$${r_g}^2 = \frac{{\sigma_{g1,g2}}^2}{\sqrt{{\sigma_{g1}}^2 * {\sigma_{g2}}^2}}$$

where N equals sample size in the discovery sample, M equals the independent number of genetic effects in the set of SNPs, $\sigma_{g1,g2}{}^2$ is the genetic covariance between target and discovery trait and $\sigma_{g1}{}^2$ equals the genetic variance explained by all measured markers in the target trait. As the discovery sample here was an ascertained case control sample (34241 cases and 45604 controls), we substituted the effective N using the effective sample size formula proposed by Wilier et al.[6]

$$N \approx \frac{4}{N_{cases}{}^{-1} + N_{controls}{}^{-1}}$$

Note that N is approximate and R^2 is estimated directly in the PRS analyses. Therefore, we needed to assume values for M and $\sigma_{g1}{}^2$. Any uncertainty in these values will not be reflected in the confidence bounds around the genetic covariance. We assumed M to equal 200.000. To explore the influence of the uncertainty in $\sigma_{g1}{}^2$ on the estimate of r_g we computed the genetic correlations assuming the heritability explained by all SNPs included in the score for childhood psychopathology to be 0.15 (Figure 3), 0.20 (Figure S1) or 0.10 (Figure S2). Note that we did not account for differences in the variance explained by the SNPs for the different psychopathologies at the different ages. We further assumed that the equations remained valid for estimates of B obtained from GEE (to correct for the presence of related samples) or ordered logistic regression (to correct for the fact that the ALSPAC phenotype was an ordered categorical variable).

Reference List

(1) Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature genetics* 2014;46(8):818-25.

(2) Abdellaoui A, Hottenga JJ, de Knijff P et al. Population structure, migration, and diversifying selection in the Netherlands. *European journal of human genetics* 2013;21(11):1277-85.

(3) Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;5(6):e1000529.

(4) Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491(7422):56-65.

(5) Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013;9(3):e1003348.

(6) Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26(17):2190-1.